1 Measures of fit and cross-validation

The cross-validation procedure used here is to divide the samples into 10 segments of almost equal size. One segment is left out and the regression parameters are found using data from the 9 segments. Then the response values of the left out segment are estimated using the regression parameters found. This is repeated for each segment. Thus at the end of the cross-validation procedure we have estimates of the response values, $\mathbf{y}_c$, such that each value in $\mathbf{y}_c$ is estimated using 90% of the data. If there are 45 samples, 4 or 5 samples would be left out at each regression. The 40 or 41 samples are used to estimate the regression parameters, which then are used to compute the response values. For each of the 10 segments the results from a=1,…,15 are registered. This is carried out L=30 times, which gives L estimates of the response values, $\mathbf{y}_{c,l}$.

The R^2 value in linear regression is defined as $R^2=1-|\mathbf{y}-\hat{\mathbf{y}}|^2/|\mathbf{y}|^2$. It has the important interpretation that it is the squared value of the simple correlation coefficient between the observed response value and the estimated response value according to the linear model, $R^2 = [(\mathbf{y}^T\hat{\mathbf{y}})/|\hat{\mathbf{y}}||\mathbf{y}|]^2$. (This follows from $(\mathbf{y}^T\hat{\mathbf{y}})=|\hat{\mathbf{y}}|^2$). In the cross-validation there is a similar situation, the response value, $\mathbf{y}$, and estimated response values from cross validation, $\mathbf{y}_c$. For comparison it is natural to define Q^2 similarly as the squared simple correlation coefficient between the response values and the estimated response values from cross validation, $Q^2= (r_{y,yc})^2$. Note that for each cross-validation 15 values of Q^2 are computed each corresponding to the dimension in the model.

2 A case study. NIR data

In Figure 3 the NIR data from measuring milk is shown. They have been measured using an instrument from Foss-Analytic (Foss). The figure shows 45 spectra. Foss prefers to calibrate instruments using relatively few spectra. NIR data show mostly smooth curves with some fluctuations at few wavelengths. Data are auto-scaled (mean centred and scaled to unit variance) before analysis.

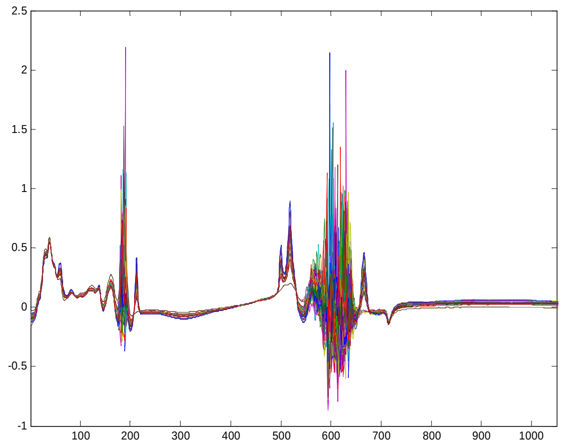


Figure 3. 45 NIR spectra from measurements of milk.

Regression using all variables

It may be instructive to look at PLS regression, where all 1056 variables are used. The instrument gives 5 response variables. Here only response variable no. 3 is used. In Figure 4 are shown the results of PLS regression using 5 components. The R^2 value is 0.9788 and Q^2=0.9668. Both the fit and the cross-validation values are fairly good, but not quite satisfactory. The yearly sales of Foss (more than 250 mio euros) are based on the best results that the mathematical models can provide with. We shall consider how these results can be improved.

Covariances and Q^2 values

In Figure 5 are shown the squared covariances, $(\mathbf{x}_i^T\mathbf{y}_3)^2$, i=1,2,…,1056. The largest value is 0.8700 which is found at variable 38. We see that there are many values close to zero. An important question is: which variables should be used? The procedure chosen here is to sort the variables according to the values of $(\mathbf{x}_i^T\mathbf{y}_3)^2$. For the first say, 10 values we get: 0.8700, 0.8659, 0.8624, 0.8542, 0.8518, 0.8496, 0.8485, 0.8471, 0.8446, 0.8419. The corresponding 10 variables are: 38, 37, 39, 36, 33, 34, 32, 35, 31, 30. The result of significance testing of section 9 is that four dimensions should be used and 105 variables should enter the model. It may be instructive to study the situation closer. For that purpose 1056 PLS regression are carried out, the first using variable 38, next using 38 and 37, and so on, where one variable is added to the model at the time. For each of the 1056 PLS regressions a cross-validation is carried out using 10 PLS regressions. Thus altogether 10560 PLS regressions have been carried out. For each of the 1056 steps the Q^2 value based on the 10 PLS regressions is computed.

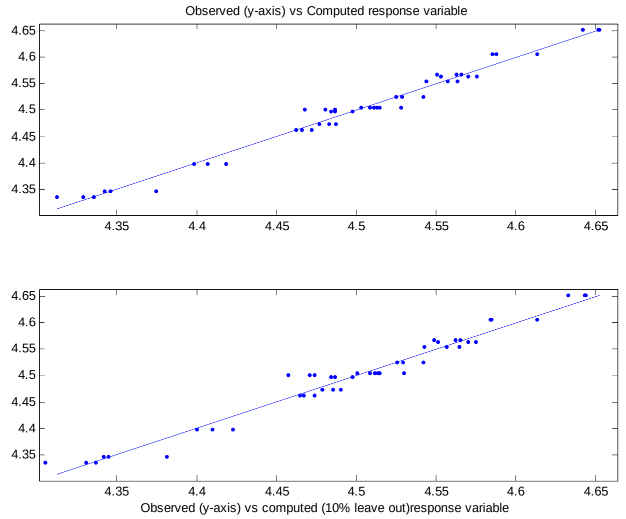
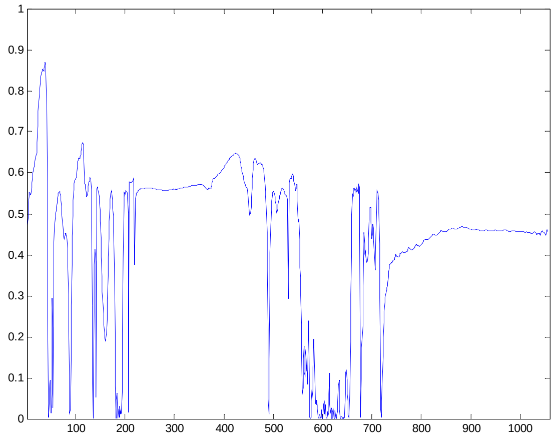


Figure 4. Observed versus computed response values. Upper one model, lower cross-validation.



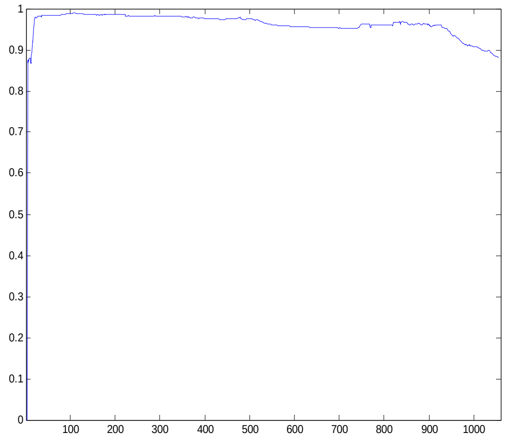


Figure 5. Plot of the values of $(\mathbf{x}_i^T\mathbf{y}_3)^2$

Figure 6. Plot of 1056 Q^2 values from PLS regressions, each based on 4 dimensions.

The resulting 1056 values of Q^2 is shown in Figure 6. The largest value of Q^2 is 0.989 and is obtained, when 105 variables are used. We typically see a curve for Q^2 that has a maximum for a certain amount of variables and then falls, when more variables are added to the model. I have seen examples, where the maximum for Q^2 is 0.95 but goes down to 0.45, when all variables are used. Therefore, it may be essential to find the appropriate variables that should be used.

The interesting issue is now: how are we doing, if only these 105 variables are used? This is shown in Figure 7. It corresponds to Figure 4. Here also 5 components are selected. The R^2 value is 0.994. Now we see that the cross-validated values also fit well to the line, Q^2=0.990. Would this be satisfactory for the company? For that purpose we show in Figure 8 the observed response values, $\mathbf{y}_3$, the residuals from the model (when 105 variables are used), $\mathbf{e}=\mathbf{y}_3-\mathbf{y}_e$, and the residuals from cross-validation, $\mathbf{e}_c=\mathbf{y}_3-\mathbf{y}_c$. We see that the residuals from the model vary within ±0.015 and the cross-validated residuals within ±0.02. This is satisfactory for Foss. It can say to its customers that the precision of the instruments is of the order ±0.02 for scaled data. This precision is also satisfactory for the customers.

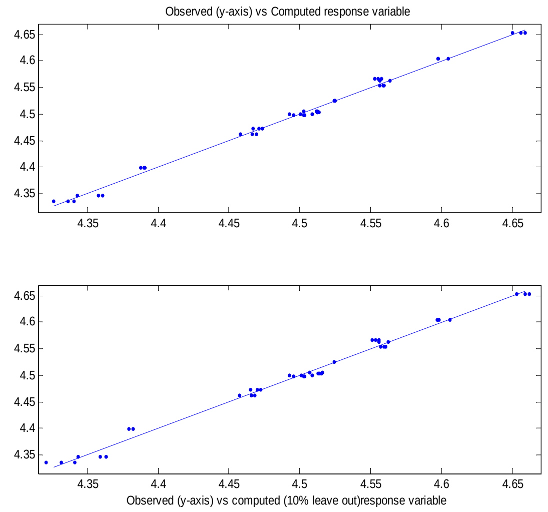


Figure 7. Observed versus computed response values. Upper one model, lower cross-validation. 105 x-variables.

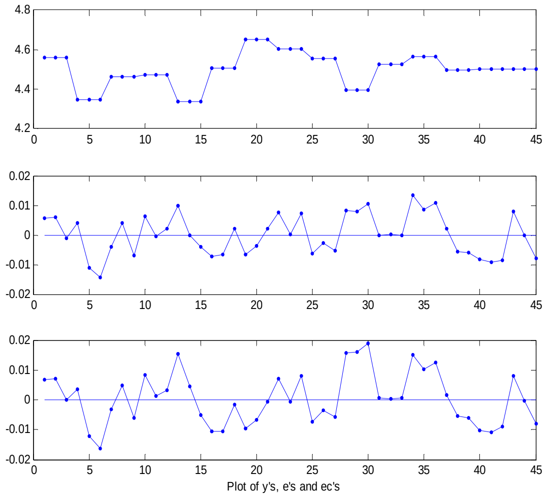


Figure 8. Plot of response values, $\mathbf{y}_3$, residuals from the model, $\mathbf{e}=\mathbf{y}_3-\mathbf{y}_e$, and residuals from cross-validation, $\mathbf{e}_c=\mathbf{y}_3-\mathbf{y}_c$.